JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECT AND PREVENT PHISING WEBSITES USING MACHINE LEARNING TECHNIQUES

[1]Sony Kurian,[2]Anita Brigit Mathew,

[1]Associate Professor Electrical Electronics Engineering Department,
[2]Associate Professor Artificial Intelligence and Data Science Department,
[1,2]Viswajyothi College of Engineering and Technology, India.

*Abstract:*PHISHING, a social engineering attack similar to cyber-crime often used to pilfer user's credentials, information, data, Bank details, passwords, etc. Example, a user can lose his/her money or valuable documents by accessing a fake website that is very much similar to the original one. There are different types of Phishing such as, spear phishing, vishing, whaling, and email phishing, etc. in this the most common form of phishing through email. Hence, to detect and stop the various cyber-crime, tremendous kinds of research are going on. In this paper we analyze, identify and classify the numerous methods used for detection of phishing.

Index Terms - Classification, Machine Learning, Feature Extraction, Phishing websites

I. INTRODUCTION

Phishing attacks are security threats where malicious actors try to get hold of user's personal information by pretending to be trusted persons. The main aim of these types of attacks is to gain important data such as credit card details, passwords, and bank account numbers. Phishing attacks are in different forms like VOIP, emails messages, tweets, and SMS. The most common is phishing through email.

Phishing usually begins with malicious attackers sending a fraudulent email to targeted victims and make them feel that these messages are authentic entities like banks, government agencies, etc. For example, there will be a Phishing website that sends emails to users which is look-alike to the original website. By this, the victims believe in the spoofed email and proceeds with the steps asked by the attacker. Normally what the attacker requested to open the specified email and fill in the details. In this way, the attacker got all the details about that person and then misuses those data. The attacker uses SMTP (Simple Mail Transfer Protocol) for fraudulent emails.

Traits or Steps in Phishing as shown in Fig. 1. The various traits include,
1.There will be a phishing website that sends phished spoofed emails to the selected users.
2.The spoofed emails usually consists of message that request action from a person.
3.When the user accesses the link in the email, it will direct to the webserver which is controlled by a phisher.
4 The user enters all the details requested by the attacker.
5.Then the user data is misused by the phisher.

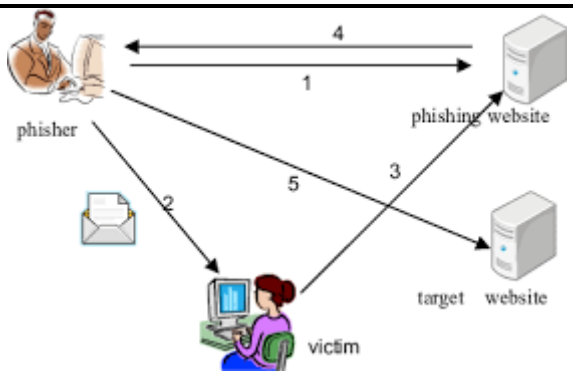

Fig. 1: Steps in Phishing Attack

There are many techniques for phishing identification and blockage which includes the use of anti-phishing freeware's, browser extension tabs, DNS and whole URL particulars, proactive detection, two reason authentication scheme, etc. We can use phishing detection as the solutions for preventing phishing and for other phishing detection solutions.

Phishing Prevention Schemes provide more security to the authentication scheme to prevent phishing attacks. These techniques are again divided as external authentication devices-based, watermark-based, image password-based, RFID-based, dynamic security model-based, QR Code based, smart device-based, and so on.

User Training Scheme educates the users via emails and another medium through that user can itself find those spoofed websites or through specific software at the host sites.

Phishing Detection Scheme detects a phishing website via web browser in client-side through specific software at the host.

II. APPROACHES

.

1. Heuristic Approach

All URL is a vector of binary features. At the time of checking, these features are fed into the web algorithm then the previously unseen URLs within the vectors are then mapped there to. Then continuous with this vector and reach the final result, as either phishing website or legitimate website.

2. Blacklist and whitelist approach

The phishing websites are identified by using machine learning and heuristics algorithms. It is a score-based mechanism, new URLs are created by heuristics and combing different divisions of the phishing websites. This is done from the selected list of blacklist websites. Then score of the URL is calculated. If the score happens to be greater than the threshold value, then the chosen website is considered to be a phishing website.

In whitelist based it is based on the online activities of the user.it based on the whitelist that is based on the user's profile which is updated dynamically whenever the user visited any website.

3. Visual Similarity Approach

As the name specified it is based on the visual similarity like detection of an image from web page similarly detection of noisy contents. This method matches the model of different visualization features like text contents such as, font colour, size, background colour fond family, etc. and text features. This is because the phisher copies the contents of the page from the actual website. The phisher may use the logos and images of real websites to target the users. There are 3 stages for it

C.1 Logo Extraction-

The website logo obtained from the sceptival website is extracted using Goldphish and it is converted into Optical Character Recognition(OCR) freeware.

C.2 Legitimate Website Extraction

The retrieved text is modelled as an input question for extraction purpose. Usually, google is pre-owned because from analysis it is found that genuine websites are returned as results through google.

C.3 Differentiations

The results obtained from search engines like google are compared with the Phishing websites. If it happens to be true that any province gets matched with the selected website, then it is declared to be a legitimate one or considered as a phishing website site.

III. LITERATURE SURVEY

Mustafa Aydin et al. [1], in their paper use classifier and data mining algorithms in order to achieve a better result. For this they have used Relief and Gain Ratio Attribute feature selection mechanisms. According to the finding analysis J48 and SMO algorithms results are satisfactory in phishing website detection. However, the Naïve Bayes algorithm also performed poorly by consuming a lot of time.

SrushtiPatil et al. [2] use five approaches namely blacklisting approach, heuristics or rule-based approach, machine learning-based approach, content-based approach, and finally the Hybrid approach, which gave better results. It gives mutual verification to the server just as the customer side. Utilizing these procedures client doesn't to reveals his credentials. It utilizes different dataset sorts based on the occasions that cover all site highlights being checked. No element ought to be left unchecked to guarantee total URL confirmation.

Moitrayee Chatterjee et al. [3]. Their model is fit for adjusting to the powerful conduct of the phishing sites and hence become familiar with the highlights related to phishing site location. The actual model is self-versatile to the adjustments in the URL structure. The issue of identifying phishing sites is a case of the traditional grouping issue. The performance of the system is estimated utilizing exactness, review, precision, and F-measure.

Mohammad Mehdi Yadollahi et al. [4] introduce a solid recognition framework which can change according to phishing sites and climate. This methodology is projected on the web and highlights a new ML method to concatenate authentic and phishing sites. Their proposed work deals with different aspects of discriminative URLs and the site page's GitHub codes. It is indeed a total arrangement from user side which doesn't require any help from outside. The exploratory outcomes feature the vigor and seriousness of our enemy of phishing framework to recognize phishing and authentic sites. The strategy utilizes various sorts of distinctive highlights, for example, highlights of URL's, HTML-XML based components, Statistical based elements, and Natural Language Processing based documents.

Yazhmozhi V.M et al. [5] uses five types of classification algorithms with two different feature sets using word vectors and natural language processing to identify the better performance one. Then the accuracy of different machine learning classification algorithms is computed and support vector machine, decision trees, logistic regression, random forest and naive Bayes algorithms are analyzed by using various features. This technique predicts tremendous phishing URL's correctly. Then it updates the URL list periodically. Similarly, it works well for visually blunt persons, and use the website for transactions.

HemaliSampat et al. [6]. Their system model spotlights on distinguishing the phishing assault dependent on checking phishing site highlights, Blacklist, and the WHOIS database. Certain features can be utilized to separate authentic and caricature pages. There are tremendous features as space character, URLs, encryption and security, page style, source code, and entities, the address bar web, and the human factor social. URL's and area names highlights are checked to utilize a few measures like long URL address, DNS, IP Address, adding a prefix or postfix, diverted image utilization.

Amani Alswailem et al. [7], discusses the framework goes about as additional usefulness to a web program as an augmentation that naturally tells the client when it identifies a phishing website. The framework depends on supervised learning. It chose the Random Forest procedure because of its great presentation in classification.

ShaheenMondal et al. [8], talks about the fundamental reason of the methodology is a crossover Machine Learning model including two stages checking with a blacklist and whitelist, and heuristics-based recognition, to expand the precision of the proposed calculation. The model is assembled utilizing Python 3.7, and SKLearn, Matplotlib, Seaborn, Numpy, BeautifulSoup, and Pandas.

IV. PROPOSED WORK

Only the blacklist and whitelist have been introduced out of all of the previous work, which has the disadvantage of not being revised in a long time. Our proposed solution is based on a hybrid approach that employs all three methods-blacklist and whitelist, heuristics, and visual similarity.

1.Create a browser extension to track all "HTTP" traffic on the end-user device. The advantage of an extension over a program or software is that the device would be more flexible.

2.Check each URL's domain against the white list of trustworthy domains as well as the blacklist of illegitimate domains. Web scraping will be used to retrieve the data required for both lists and stored in the control unit called server. Suppose, if the URL's province is found to be in the white-list label then its considered to be innocent else we use other approaches to find the result.

3. Besides, the review from entire website, we can now take review performed by taking into account different data (feature extraction). We looked at the following characteristics: website protocol (secure or insecure), URL length, hyphens (-) number in the URL, list of @ symbols in the URL, and data of dots in the URL. This is done using the Alexa ratings, or direct IP address, etc.

- If the number of hyphens in the URL is greater than one, the website is phished.
- If the hyphen in the URL equals 1, the website is suspect.
- If the URL contains a hyphen, it is a legitimate website.

4. Intuitively, the greater the resemblance between the target and phishing page, the more likely the users would be duped. Assailants, this is the excuse.

5. To counteract such conduct, we'll take suspicious URL's and analyze the CSS of every unauthorized URL. Hence, this methodology will help us to investigate visual aspects.

6. The collected data from the previous step will be run via machine learning classifiers such as Regression Trees, Decision Trees, and logistics.

7. The match and similarity scores are determined. If the score exceeds the threshold, the score is marked as phishing website and then block it.

8. Thus, the above method essentially helps in a three-level security lock on the phishing website, making it more effective and precise than any other device currently in use.

V. IMPLEMENTATION

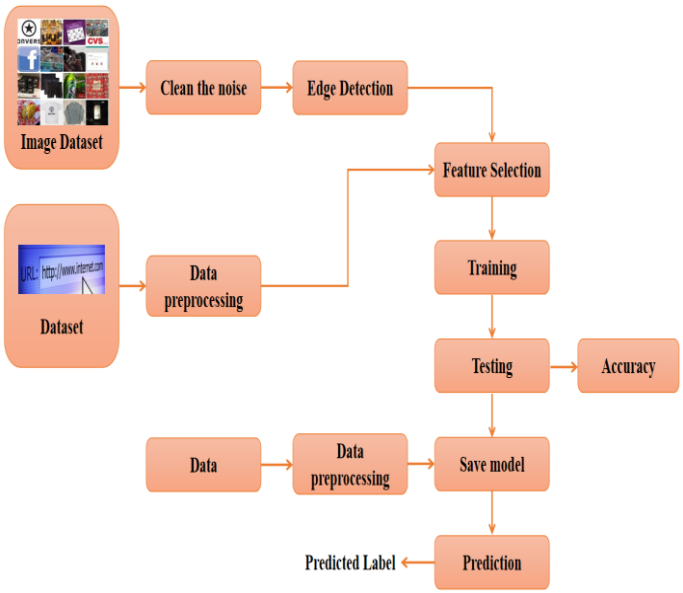

Fig. 2: System Model for the detection and prevention of Phishing Attack

Step 1: DATASET

Dataset can be defined as a collection of information. These data sets can be called as training datasets which is used to training a model. An example of the dataset is a table with rows and columns and column headers and each column and row represent a particular data as shown in Fig. 2.

Data set can be defined as a backbone for training algorithms without a good dataset Artificial Intelligence would be difficult to understand because machine learning is based on this dataset. without a good dataset, we can't achieve success in the project.in this project, we are using 3 datasets for preparation, testing, training, and model selection.

Step 2: CLEAN THE NOISE

When there is noise present in the machine learning algorithm it will cause problems because if we not trained them correctly the algorithm can start generalizing from it. So here we are using a process called image denoising which is for removing noise from an image.We perform this image denoising to restore the image. image noise can be defined as change or variation in properties of an image like brightness, color, etc. The image noise can be produced by an image sensor, scanner, digital camera, etc. Image denoising can be very useful in many aspects like image restoration, image registration, visual tracking, image segmentation, image classification wherein all these obtain an image without noise. Hence, we can say it plays an important role in evaluating the accuracy. There are various methods for image denoising like total variation regularization, local regularization, sparse representation, low and minimization, etc.

Step 3: EDGE DETECTION

Edge detection is a basic image processing technique that entails computing an image gradient in order to measure the magnitude and direction of edges in an image. Line detection, feature detection, and image classification are only a few of the downstream tasks in computer vision that use image gradients.

Image processing relies heavily on edge detection. Despite developments in deep-learning-based techniques such as Convolutional Neural Networks, which can detect very complex edges (i.e. edges with varying curvature, noise, color, and so on), classical edge detection remains a challenge.If the data is considered to be easy and predictable, for example, a Canny Edge Detector would operate right out of the box, while a CNN would be more difficult to implement.

Step 4: DATA PREPROCESSING

Data preprocessing can be defined the initial step in machine learning model creation, it is the most important part of the model creation that increases the quality of the data. It can be defined as a process of converting raw data with lots of errors into data that is readable and understandable. The raw data consists of lots of errors like noise, missing values, etc and the data preprocessing can be used to remove all these, format, and organize and this can be used for training the classifier model which increases the overall performance of the system.

Step 5: TRAINING
As the name specified training is the process of training the algorithm or machine learning model and to perform this we use a data known as training data. A training model consists of input data and corresponding output data and it is the dataset we are used to train the ML model. and this is known as supervised learning and unsupervised learning means allowing the algorithm to work on its own, here we no need to supervise the model.

Step 6: FEATURE SELECTION
Feature selection is defined as the process of selecting features for model creation in machine learning. The accuracy and speed of training data can be increased and speed up by limiting the number of features we use. One of the features of feature selection is that data contain some irreverent features that can be removed without losing information. New features are created by feature extraction from original features where feature selection returns the subset of features.

Step 7: TESTING
Testing is defined as evaluating the output of the model in terms of accuracy and precision to check the accuracy of the model. By testing a model, we are assured of the quality of the system. The overall performance of the design can be checked by two training datasets, a validation dataset, and a test dataset.

Step 8: SAVE MODEL
In machine learning, we need to save the trained models in a file. By this, we can restore them when we need them and can compare them with other models and test them on new data. The process of saving data is known as serialization and the restoring of data is defined as Deserialization.

Step 9: ACCURACY
Accuracy can be defining as the performance measuring technique. Accuracy is the correct predictions percentage and it can be calculated by dividing the total count of corrected predictions by the count of all predictions. Improving model accuracy lowers the cost even though errors have a high cost. For example, in case of cancer diagnosis, a false-positive costs both patient and hospital. The advantages of increasing system accuracy include benefiting time, money, and excessive stress.

Step 10: DATA
The testing set is a collection of observations used to assess the model's success using various performance metrics. The evaluation set does not contain any findings from the training set. It would be difficult to tell whether the algorithm has learned to generalize from the training set or has merely memorized it if the testing set contains examples from the training set.A software that generalizes well would be able to perform tasks with new data effectively. A program that memorizes the training data by studying an overly complex model, on the other hand, maybe able to correctly predict the values of the answer variable for the training set, but not for new instances. Over-fitting is the practice of memorizing the training package. A software that memorizes its observations will not perform well because it may memorize noise or coincidental connections and structures.

Step 11: PREDICTION
Predict means predicting the output of a model. it can be defined as the final result that obtained after we trained a model with a dataset. Prediction can be performed during the model development process, after the model has been created, or after a failure has occurred. Predictive modeling refers to the process of creating a model that can make predictions. These predictions are done by using machine learning algorithms

VI. CONCLUSION

The future system enables the users for safe browsing and careful transactions. It helps users to keep their private information without the fear of losing it. As long as our anticipated system to users in the kind of augmentation makes the means of delivering our system greatly easier. To happen as expected in this context, the algorithms that recurrently adapt to new examples and skin of phishing URLs. This new model provides highest accuracy to the users. A point of challenge in this domain is that criminals are constantly creating new strategies to counter our protection measures. To happen as expected in this context, we penury algorithms that recurrently adapt to new examples and skin of phishing URLs. And like so we help online knowledge algorithms. This new system

container is considered to purpose the highest accuracy. utilizing another approach when all's said and done will enhance the accurateness of the system if a cost-effective system.

REFERENCES

[1] Mustafa Aydin, Ismail Butun, Kemal Bicakci, Nazife Baykal, "Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs", IEEE Xplore, May 30, pp. 34-41, 2020.

[2] SrushtiPatil, SudhirDhage, "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework", 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 10-18, 2019

[3] Moitrayee Chatterjee, Akbar SiamiNamin, "Detecting Phishing Websites through Deep Reinforcement Learning", IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 227-232, 2019.

[4] Mohammad Mehdi Yadollahi, FarzanehShoeleh, ElhamSerkani, AfsanehMadani, Hossein Gharaee, "An Adaptive Machine Learning-Based Approach for Phishing Detection Using Hybrid Features", 5th International Conference on Web Research (ICWR), pp. 281-286, 2019

[5] Yazhmozhi. V.M, Dr. B. Janet, "Natural language processing and Machine learning-based phishing website detection system", Third International Conference on I-SMAC, IEEE Xplore, pp. 336-340, 2019.

[6] HemaliSampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes, "Detection of Phishing Website Using Machine Learning", International Research Journal of Engineering and Technology (IRJET) Vol. 5, Issue: 03, pp. 286-291, Mar-2018.

[7] Amani Alswailem, Bashar Abdullah, Norah Alrumayh, Dr.AramAlsedrani, "Detecting Phishing Websites Using Machine Learning", IEEE Xplore, vol. 2, pp. 78-83, 2019.

[8] ShaheenMondal, DikshaMaheshwari, NilimaPai, AmeyaaBiwalkar, "A Review on Detecting Phishing URLs using Clustering Algorithms", IEEE Xplore, vol. 5, pp. 89-94, 2019.